14th International Conference on Information Fusion
Chicago, Illinois, USA, July 5-8, 2011

Fusion of Disparate Information Through Joint Embeddings

David J. Marchette* Jeffrey L. Solka*

Abstract

Most pattern recognition tasks can be abstracted to a problem of utilizing comparisons between objects to perform the given inference task. Often these comparisons are in the form of a distance measure or dissimilarity. The design of appropriate comparison functions for particular inference tasks is an area of extensive research, and often rests on expert knowledge of the problem domain. If the data of interest come from two different sensors, or consist of very different types of data, a single dissimilarity may be inappropriate; instead, one might utilize several dissimilarities, each designed for a specific sensor or data stream. In this work we consider the problem of fusing information obtained from very different sensors or sources, encoded through the use of dissimilarity functions. Given n observations from source j, we have an $n \times n$ dissimilarity measure D_j, and we wish to utilize all this information in our inference. We describe several methods of utilizing these dissimilarity matrices that are based on embedding the observations into a single space. These methods optimize either the fidelity (whether the distances in the embedded space match the original dissimilarities) or the commensurability (whether matched objects from different sensors are close in the embedded space) or both. We discuss the properties of these embeddings, apply the idea to a problem in network modeling, and point out some interesting areas of further research.

1 Introduction

Consider the problem of fusing the information from two sensors in order to perform a given inference. In the case we consider in this paper, there will be two different sets of observations from two different sensors, and we wish to combine the observations from the two sensors. We will consider only the case of two sensors, but the multiple sensor case can be analyzed similarly. Much of the work in this paper has been reported in a number of papers, in particular Priebe et al. [2010], Marchette [to appear] although we will consider some new results and provide some new insights into the methodologies.

*Naval Surface Warfare Center, Code Q33

978-0-9824438-3-5 ©2011 ISIF

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **JUL 2011**	2. REPORT TYPE	3. DATES COVERED **00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE **Fusion of Disparate Information Through Joint Embeddings**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Surface Warfare Center,Code Q33,1333 Isaac Hull Ave, SE ,Washington Navy Yard,DC,20376-7101**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the 14th International Conference on Information Fusion held in Chicago, IL on 5-8 July 2011. Sponsored in part by Office of Naval Research and U.S. Army Research Laboratory.

14. ABSTRACT

Most pattern recognition tasks can be abstracted to a problem of uti- lizing comparisons between objects to perform the given inference task. Often these comparisons are in the form of a distance measure or dis- similarity. The design of appropriate comparison functions for particular inference tasks is an area of extensive research, and often rests on ex- pert knowledge of the problem domain. If the data of interest come from two different sensors, or consist of very different types of data, a single dissimilarity may be inappropriate; instead, one might utilize several dis- similarities, each designed for a specific sensor or data stream. In this work we consider the problem of fusing information obtained from very different sensors or sources, encoded through the use of dissimilarity func- tions. Given n observations from source j, we have an n n dissimilarity measure Dj , and we wish to utilize all this information in our inference. We describe several methods of utilizing these dissimilarity matrices that are based on embedding the observations into a single space. These meth- ods optimize either the fidelity (whether the distances in the embedded space match the original dissimilarities) or the commensurability (whether matched objects from different sensors are close in the embedded space) or both. We discuss the properties of these embeddings, apply the idea to a problem in network modeling, and point out some interesting areas of further research.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Formally, let Ξ be a space, and $\pi_0 : \Xi \to \Xi_0, \pi_1 : \Xi \to \Xi_1$ continuous maps into two dissimilarity spaces. Recall that a dissimilarity space X is a space in which a function $d : X \times X \to \mathbb{R}$ is defined with the properties: 1) $d(x, y) \geq 0$; 2) $d(x, y) = 0 \iff x = y$. The maps π_i can be considered to be the measurements from two sensors. Let ρ_i be embeddings defined on Ξ_i into a space X (in this discussion $\mathbb{R}^d$ for a fixed d). Throughout this paper we will assume the embeddings are performed through multidimensional scaling (MDS) (Borg and Groenen [1997]) without explicitly defining which specific approach to MDS is used.

$$\begin{tikzcd}
 & \Xi \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_0"] & \\
\Xi_1 \arrow[dr, "\rho_1"'] & & \Xi_0 \arrow[ll, dotted, "\varphi"'] \arrow[dl, "\rho_0"] \\
 & X = \mathbb{R}^d &
\end{tikzcd}$$

Let $X_n = \{x_1 \ldots, x_n\} \subset \Xi$ be observations in the original space, and denote by x_i^j the image of x_i under π_j: $x_i^j = \pi_j(x_i)$. φ is an unknown (and possibly fictitious) "manifold matching" function. Note that we assume we have no way to observe Ξ directly, we can only observe the image under the π_i. We also assume that we have no idea what the "manifold matching" function φ is, and in fact in many instances of interest, this function may not even exist – it is only notional. Thus, this work is not aimed at trying to estimate φ (although, see Marchette [to appear]).

Our methodologies will all start with dissimilarities d_i defined on Ξ_i. We will denote by Δ^i the $n \times n$ inter point dissimilarity matrix on the Ξ_i portion of $\widehat{X}_n$. An obvious approach is to embed each sensor into a separate space, and then form the product: essentially forming the union of the features from the two sensors. This makes no use of any redundancy or correlation of the information in the two sensors; however, when the sensors are sufficiently different then an approach like this might well be what is required. We will discuss below some ways of investigating whether this might be the case.

2 A Separate Embedding Approach

$$\begin{tikzcd}
 & \Xi \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_0"] & \\
\Xi_1 \arrow[d, "\rho_1"] & & \Xi_0 \arrow[d, "\rho_0"'] \\
\mathbb{R}^d & & \mathbb{R}^d \arrow[ll, dotted, "Q"']
\end{tikzcd}$$

The idea of this is to define the embeddings ρ_i independently, then define a mapping Q that maps the x_j^0 as close as possible to the matching x_j^1. This is usually performed by Procrustes:[1]

$$\arg\min_{Q^TQ=I} \|\widetilde{X}^1 - \widetilde{X}^0 Q\|.$$

We define the fidelity of the embedding in terms of raw stress (see Borg and Groenen [1997] for discussion of this and other criteria that might be used in its place). The fidelity measures how well the embedded points match their respective dissimilarities:

$$\mathcal{F}^k = \sum_{i=1}^{n-1} \sum_{j>i} (\Delta^k_{ij} - d(\widetilde{x}^k_i, \widetilde{x}^k_j))^2.$$

It is well known that MDS minimizes the fidelity error.[2] Thus, the separate embedding approach optimizes the fidelity of each embedding separately.This does not guarantee that the resulting two point sets are commensurate. The Procrustes embedding is designed to optimize this commensurability, under the rigid motion constraint, which ensures that the fidelity is retained. We define the commensurability error as:

$$\mathcal{C} = \sum_{i=1}^{n} (d(\widetilde{x}^0_i, \widetilde{x}^1_i))^2.$$

In this definition we abuse notation by using the same symbol $\widetilde{x}^0_i$ to denote the image of $\widetilde{x}^0_i$ under the Procrustes transformation. This allows us to refer to the commensurability as a criterion on the ultimate embedded points regardless of how the embedding is performed.

Similarly, a canonical correlation approach can be used to optimize commensurability without regard to fidelity. For the details see Priebe et al. [2010].

3 Joint Embedding

Define $\Delta^\lambda = \lambda\Delta^1 + (1-\lambda)\Delta^0$. The joint embedding approach we consider utilizes the three inter point dissimilarity matrices, Δ^1, Δ^0 and Δ^λ for a fixed $\lambda \in [0,1]$ (usually $\lambda = \frac{1}{2}$). Form the $2n \times 2n$ omnibus dissimilarity matrix:

$$\Delta = \begin{pmatrix} \Delta^1 & \Delta^\lambda \\ \Delta^\lambda & \Delta^0 \end{pmatrix}.$$

Then use Δ to define the embeddings.

[1] For simplicity we are assuming that the dissimilarities are on the same scale, so that no scaling is required of the Procrustes transformation. We also assume that they are centered, so that we do not have to translate (although both of these can easily be incorporated in the Procrustes methodology).

[2] Different methods of MDS have been developed for minimizing various criteria. As discussed above we focus on raw stress for ease of exposition.

It should be noted that this approach assumes that the dissimilarity matrices are on the same scale. In the separate embedding approach one can incorporate scale into the Procrustes transformation, but the joint embedding method should be used on matrices that have been scaled appropriately. How best to do this is a topic for another time.

We are optimizing:

$$\mathcal{E} = \sum (d(X_i^0, X_j^0) - \Delta_{ij}^0)^2 + \sum (d(X_i^1, X_j^1) - \Delta_{ij}^1)^2 + \sum (d(X_i^0, X_j^1) - W)^2.$$

In this, W is taking the place of $d(X^0, X^1)$ which is unknown, and in some applications, may not even be meaningful. Note that we impute the diagonal of W to be 0. Given our notion that matched documents are "the same", it is reasonable that their distance should be small, and this choice attempts to force this. However, we know experimentally that this is not the optimal choice for the diagonal of W. Clearly the choice of this matrix is an area for future research.

The diagonal of the third term is the commensurability error. We call the off-diagonal term the separability error. That is, we also want to ensure that non-matched pairs are appropriately far apart.

4 How Can This Fail?

Jointly optimizing fidelity and commensurability seems to be an excellent strategy, but one might wonder if there are cases in which it cannot be effective. In Priebe et al. [2010] is a brief discussion of this in terms of Hausdorff distance. Consider the case where the spaces are Euclidean and the embeddings π_i are linear projections onto subspaces. The Hausdorff distance between two subspaces is $2\sin(\theta/2)$, where θ is the canonical angle between the subspaces. Essentially, this says that if the subspaces are too far apart – orthogonal – the points will not be commensurate. To turn this around, highly incommensurate points in the training data are indicators that the embedding approach proposed may be inappropriate.

5 Experimental Results

We apply the approach, suitably modified, to a problem in modeling random graphs. The model we investigate is the random dot product graph (RDPG) model (see Marchette and Priebe [2008]), in which each actor in a social network has a vector of attributes, and the edge probabilities are a function of the dot product of these vectors. In this problem we have external measurements related to these vectors, and wish to use this extra information to improve the model fit. In this case the joint embedding approach is very well suited, and produces a gratifyingly large improvement in fit.

The random dot product graph (RDPG) model is a simple model of social networks that relates attributes of the actors (vertices) to their social relationships (edges). For each vertex v is given a random vector X_v, and the probability of an edge between u and v is given by:

$$P[u \sim v] = X_u^T X_v.$$

The edges are conditionally independent given the X's. Given a graph G on n vertices (all graphs will be simple (no self-loops or multiple edges) and undirected), we wish to fit the model: find $\widehat{X}$ that "best" fits the graph. We will use the convention that X and its estimates are $n \times d$ dimensional matrices.

We define *best* in terms of the Frobenius norm: given the adjacency matrix A of the graph, minimize

$$\|A - \widehat{X}\widehat{X}^T\|_2^2.$$

Thus, we are considering squared error as our criterion. Note that we can solve this easily via spectral methods. The optimal solution is available through the eigenvalues and eigenvectors of A. Note further however that this is not quite what we want since the diagonal of the adjacency matrix should be ignored. We thus augment the diagonal with and estimate of the norm of the X_v: it can be shown that the expected degree of a vertex $Ed_v = (n-1)E\|X_v\|$ (this is exactly analogous to the similar formula for Erdös-Renyí graphs). Define

$$\breve{A} = A + \text{diag}(\frac{d_v}{n-1}),$$

and now minimize

$$\|\breve{A} - \widehat{X}\widehat{X}^T\|_2^2.$$

The optimum is found as $\widehat{X} = \sqrt{\Lambda}U$, where Λ is the diagonal matrix containing the d largest eigenvalues of $\breve{A}$ and U is the $n \times d$ matrix formed from the corresponding eigenvectors.

Now, suppose we observe attributes Y for the vertices that are correlated with the model parameters X. Can we use these to obtain an improved fit to the graph? The answer is investigated in the next set of experiments. In all experiments, $n = 100$ (the number of vertices in the graph) and we run each experiment $N = 100$ times to obtain the box-plots in the Figures below.

Given a $d + 1$-dimensional vector X, define X^- as the d-dimensional vector consisting of the first d components of X. In our experiments we will take $d = 3$. Let $p_1 = (0.0, 0.4, 0.1, 0.5)$ and $p_2 < -(0.6, 0.1, 0.1, 0.1)$, then in our experiments half of the vertices are distributed $X \sim \text{Dirichlet}(100\mathbf{p_1} + 1)^-$, and half distributed $X \sim \text{Dirichlet}(100\mathbf{p_2} + 1)^-$. These result in small clouds of points around the centering points p_i. Let $G \sim \text{RDPG}(X)$ denote the RDPG model defined by X. We will consider two choices for Y in the following experiments: 1) $Y \sim \text{Dirichlet}(rX + \mathbf{1})$ (here r is a parameter corresponding roughly to the inverse of the variance), 2) $Y \sim \text{Dirichlet}(rX + \mathbf{1})^-$. So in the second, the dimension of Y is $d' = 2$. In all experiments, we use as error the difference between the estimated probability and the true: Error $= \|XX^T - ZZ^T\|_2^2$ for estimate Z.

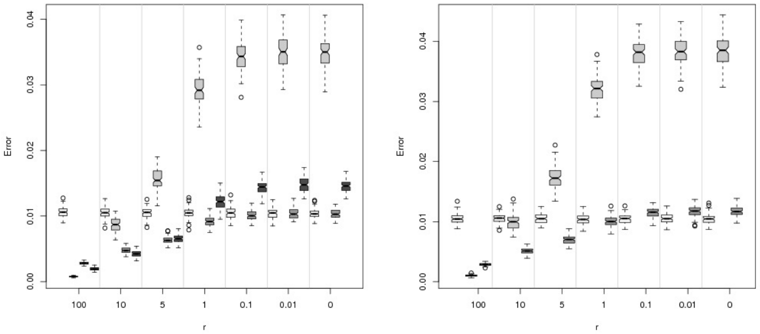

Figure 1: RDPG experiments. In increasing darkness of the boxes, these correspond to the original estimate $\widehat{X}$ from the graph alone; the estimate given by $\widetilde{Y}$ alone; the fusion result; the estimate given by the average of $\widehat{X}$ and $Q\widetilde{Y}$, with Q the Procrustes transformation to map these together. In the left plot, $d = d' = 3$ and in the right $d' = 2$.

Denote by $\widetilde{Y}$ the observed value of Y. Given an estimate of $\widehat{X}$ (say using the above spectral algorithm) we can use the Procrustes transform to define Q to map $\widehat{X}$ and $\widetilde{Y}$ together, giving one estimate as the average: $(\widehat{X} + Q\widetilde{Y})/2$. $\widetilde{Y}$ gives us a third estimate (ignoring the graph altogether), and a forth is through the joint embedding as follows. Let $B = \widetilde{Y}\widetilde{Y}^T$, $W = (checkA + B)/2$, and form

$$\Sigma = \begin{pmatrix} \check{A} & W \\ W & B \end{pmatrix}.$$

Treat this as in the spectral algorithm above, obtaining a $2n \times d$ dimensional matrix, and return the average of the top n vectors with the bottom n vectors (pairing the ith vector with the $(n + i)$th). For case 1 above, the results are shown in Figure 1 (left), and for case 2, in Figure 1 (right). As r increases, Y_v has less and less variance, and as r decreases Y_v increases variance, until at $r = 0$ Y is uniform in the simplex independent of the corresponding X value. Note that this approach, while not a dissimilarity approach, is actually quite similar to the joint embedding approach we discussed above. Classical multidimensional scaling essentially performs the spectral estimate on a matrix related to Σ above – a centering of the squared adjacency matrix that moves from dissimilarity space into dot product space, if you will. Thus, we view this approach to RDPG as all part of the same set of algorithms.

Note that for case 1, a simple averaging of $\widehat{X}$ and $\widetilde{Y}$ is possible (assuming that we have performed a Procrustes to ensure that these are commensurate) and this is depicted in the figure as the darkest box. The notches on the boxes give an idea of significance: if the notches overlap, the two approaches cannot be considered significantly different. For this particular simulation (and our goal of investigating the properties of our method) the case $r = 1$ is the sweet spot:

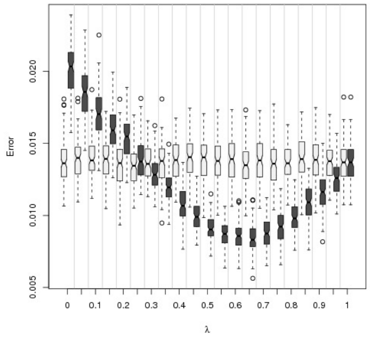

Figure 2: Third RDPG experiment. Comparison of the estimate of X using the graph only (light gray) and the estimate obtained by adding noise to the graph. See the text for details.

the joint embedding approach is the best by far, and yet the noise on the $\widetilde{Y}$ is so extreme that it is itself a terrible estimate of the model's vectors.

In the second experiment, we do not have the option of averaging (since we assume we do not know what combination of the three coordinates of X we are measuring with Y) and so we only plot the results for the two individual approaches and the joint embedding. The results are given in Figure 1 (right). Clearly the combination of the information is superior to the separate ones.

Note that the joint embedding approach is relatively insensitive to noise – in fact it seems to ignore the attributes for $r = 0$. Adding noise can (somewhat counter-intuitively) sometimes improve estimates, and to some degree that is what is happening in the $r = 1$ case.

To better understand this phenomenon, consider Figure 2. We perform the following experiment: Let both X and Y be drawn uniformly in the simplex, independently. Thus there is no "signal" in Y. Form the RDPG graph G from X and let A be its adjacency matrix as above. For $\lambda \in [0,1]$, form $B^{\lambda} = \lambda\breve{A} + (1-\lambda)\widetilde{Y}\widetilde{Y}^T$. Then use the spectral approach on B to obtain the estimate $\widehat{X}^{\lambda}$. We plot the errors of this approach compared to the estimate which uses $\breve{A}$ only (as in the previous plots) in Figure 2. Note first that, as expected, when $\lambda = 0$ the estimate is bad (there is no information about the graph in B^{λ} and when $\lambda = 1$ the estimates agree perfectly. The interesting part occurs in the range $\lambda \in [0.3, 0.95]$. As long as there is some information from the graph, averaging in this noise improves the estimate. Whether this is because the binary representation of the probabilities from the adjacency matrix does not give the algorithm sufficient flexibility to fit the model (all those 0 probability estimates that should not be 0) or for some other reason is an area for future research.

6 Discussion

We have described a method for combining information from two sensors, and the method easily extends to multiple sensors. We discussed the situations in which the method does not work, at least not without some modifications or significant work. This discussion links a measure of comparability of the sensors – the commensurability of the embedded points – to the ultimate performance of the inference, and thus provides a useful method for diagnosing when the algorithm is likely to be applicable and when not.

The joint embedding method is clearly worth considering when the data have a natural dissimilarity function available, or when the data come in the form of a dissimilarity matrix (which is often the case in Psychological experiments and in some Brain mapping experiments). When the data are presented as features, other methods of fusion immediately suggest themselves and should not be ignored. Simply forming the product (appending all the features together into one long vector) and then performing feature selection and dimensionality reduction is a well-used and time-honored approach, and should always be a part of our toolkit.

The random graph experiment showed some surprising results. Adding noise to the graph can improve estimation using the spectral approach, and the joint embedding provides a natural way to take advantage of this. Further research is clearly suggested, and this will be one of the major areas in which we will be involved in the future.

7 Acknowledgments

This work was performed under the Air Force Office of Scientific Research (AFOSR) grant NCRADA-NSWCDD-11-115 and the In-House Laboratory Independent Research (ILIR) program of the Office of Naval Research (ONR).

References

Ingwer Borg and Patrick Groenen. *Modern Multidimensional Scaling.* Springer, New York, 1997.

David J. Marchette. Implicit translation. *Wiley Interdisciplinary Reviews: Computational Statistics*, to appear.

David J. Marchette and Carey E. Priebe. Predicting unobserved links in incompletely observed networks. *Computational Statistics and Data Analysis*, 52: 1373–1386, 2008.

Carey E. Priebe, David J. Marchette, Zhiliang Ma, and Sancar Adali. Manifold matching: Joint optimization of fidelity and commensurability. *submitted for publication*, 2010.